Cuing with word senses: A test of generation-recognition theory

MICHAEL J. WATKINS and NORMAN W. PARK
University of Toronto, Toronto, Ontario, Canada M5S 1A1

It has been argued that recent demonstrations of recognition failure of words subsequently recollected contradict the generation-recognition theory of recall. This argument has been countered with the proposal that the various senses of a word are represented in the associative network by separate nodes, so that recognition may fail because of access to the wrong node. A prediction of this modified version of generation-recognition theory is that the probability of recognizing a word can be increased by providing cues to its various senses. Two experiments are described in which comprehensive definitions of the target words were given during a recognition test. These definitions failed to improve recognition, and hence this modified version of generation-recognition theory was not supported.

According to one currently popular account of episodic memory, an encounter with an item results in the tagging of some permanent representation of that item. Thus, if a randomly selected word is presented to a subject in a memory experiment, an occurrence tag is assumed to be attached to the node that corresponds to that word in a hypothetical associative network.

The success of this *tagging theory* owes much to its straightforward account of recall and recognition. Recall is assumed to entail two processes: (1) a search for, or generation of, candidate nodes in the associative network, and (2) a decision about the presence or absence of an appropriate tag. Recognition differs from recall in that only the second of these processes is involved, for the test item is assumed to guarantee access to the appropriate node just as in the case of the original encounter. Because the recognition process is a component of the recall process, it follows that a recallable item will necessarily be recognizable.

This prediction is contradicted by research of Tulving and his associates (e.g., Tulving & Thomson, 1973; Watkins & Tulving, 1975) demonstrating recognition failure for recallable words. This research usually entails the interpolation of a recognition test between the study and test phases of a paired associate procedure. The recognition test includes the paired associate response terms along with new (lure) items. Typically, the response terms are recognized with a lower probability than they are subsequently recalled.

The finding of recognition failure of recallable words has forced a revision of tagging theory, such that most common words are now assumed to be represented in the associative network by not one node but several, one for each sense of the word (Martin, 1975; Reder,

Anderson, & Bjork, 1974). Thus, if grasp-BABY is presented in the study list, only one (or at most a subset) of the nodes is tagged. When BABY is later encountered as a recognition test item, access may be made to a different, and therefore untagged, sense node, in which case recognition will fail. If the subsequent presentation of the stimulus term is assumed to result in more reliable access to the tagged node, then the level of recall will be higher than that for recognition.

This modification of tagging theory has meant sacrificing specification of the node(s) used in any given situation. Hence, the theory is stripped of much of its predictive powers. Nevertheless, the theory still retains some predictive power and is therefore testable. For instance, it predicts that recognition should improve if the nominal item is accompanied by information that is likely to increase the probability of access to the tagged node.

Our inability to specify the tagged node should not mean that we cannot increase the probability of access to it. In general, access to a given node can presumably be achieved by providing cues for the sense represented by the node; indeed, this is the mechanism that modified tagging theory assumes to underlie recall of unrecognized words in the Tulving and Thomson (1973) procedure. Thus, the probability of access to a tagged node in the recognition test, and hence the probability of recognition, can be increased by supplementing the test item with a hint about the relevant sense of the word. The difficulty of not knowing which sense is relevant can be overcome by providing as many different senses of the word as possible; a list of at least many of the shades of meanings of a word is readily available in the dictionary. In short, providing the subject with a word's senses and subsenses as listed in the dictionary should reduce his difficulties in finding the various nodes that correspond to the word, and therefore increase the probability of access to any given node, including the target (tagged) node. The higher rate of

access to the target node should be reflected in a higher level of recognition.

This prediction does not assume a direct correspondence between the senses of a word listed in the dictionary and those represented as nodes in the subject's memory. It assumes only that at least some of the dictionary senses overlap with at least some of the memory senses. These dictionary senses should then facilitate recognition just as input cues are assumed to facilitate recall.

We now describe two basically similar experiments testing this prediction. These experiments use Tulving and Thomson's (1973) materials and follow closely their method in all except the recognition test. The recognition test items were presented one at a time. For each item the subject first made a recognition judgment, next read the dictionary entry for the item, and then gave a second recognition judgment. The question of interest was whether recognition would improve from the first to the second judgment.

EXPERIMENT 1

Method

Materials and design. Two variants of the critical list were used; half of the subjects saw one, and half the other. These lists were the ones used by Tulving and Thomson (1973, Table 1). They comprised 24 weakly associated cue-target pairs, such as *ground*-COLD, *head*-LIGHT, *bath*-NEED. Six further cue-target pairs were taken from Tulving and Thomson's set lists to form a practice list.

Lures were used in both the recognition and recall tests. For each subject eight new cue-target pairs were selected from what was for him the nonpresented variant of the critical list; there was a separate random selection of these pairs for each subject. The target members of these pairs served as lures in the recognition test, the cue members as lures in the recall test. In both recognition and recall, the order of presentation of the test items was separately randomized for each subject.

The definitions of the test items, used in the recognition phase, were taken from Webster's *New Collegiate Dictionary* (1973). The length of the definitions ranged from about 30 to about 960 words.

Subjects and procedure. Sixteen volunteer introductory psychology students were tested individually or in pairs. First, a short practice trial was given. The subjects were told that they would see a series of word pairs, of which the "target" word was in uppercase and the "cue" word in lowercase. The practice list comprised six word pairs presented by a projector at a rate of 3 sec per word pair. Immediately following presentation, the subjects were given the list cues in a new, random order and they attempted to recall the corresponding target words.

The critical list, comprising 24 word pairs, was then presented and subsequently tested with procedural details the same as those for the practice list. But unlike the practice trial, two additional tasks were interpolated between list presentation and the recall test. The first was a free-association task, included merely to reduce memory for the target items to a level sensitive to experimental manipulation. The second interpolated task was the recognition test. There were 32 test items: the 24 targets and 8 new words or lures. The test items were typed on flash cards and shown one at a time. On the presentation of each card, the subject first indicated on his response sheet whether he thought the item had appeared in the critical presentation list or was a lure, and he added a confidence rating of

his judgment using a 4-point scale. He was then given a card containing a copy of the dictionary entry for the test word, and was instructed to read it carefully "because it might improve your recognition performance." After reading the definition, the subject gave a second recognition judgment and confidence rating. This sequence of recognition rating, definition reading, and recognition rating was repeated for each of the test items. The subject proceeded through the recognition test at his own pace, taking on average about 45 min.

Finally the recall test was given. Here the subject received a list of 32 items, of which 24 had been shown as list cues and 8 were lures; the lures were included to provide an estimate of target guessing. The subjects recalled the targets at their own pace.

Results

The question of critical interest is this: Did reading the dictionary definitions prompt recognition of more target items? The data provide a straightforward answer: No. In their first recognition judgments the subjects designated a mean of 16.56 of the 24 target items as targets; in their second judgments this mean was no higher—in fact, it *dropped* to 14.87. Moreover, of a total of 184 targets for which the recognition rating changed from the first to the second judgment, only 55 changed in the target direction, compared to 129 that changed in the lure direction. Separate analyses on the subset of items subsequently recalled revealed the same slight drop in performance (both in terms of hit rate and confidence rating) observed for the unconditional analyses.[1]

In the recall test 39% of the targets were recalled, and no one guessed the designated target of a lure cue. That recall was somewhat lower than that usually found by Tulving and his associates probably reflects the protraction of the recognition test. If only those words tested in the last quartile of the recognition test are considered, then recall (40%) is higher than (corrected) recognition (32%).

EXPERIMENT 2

The second experiment differed from the first in two major respects: (1) The recognition judgments, both those before and those after study of the dictionary entries, involved a two-alternative forced-choice procedure instead of a free-choice procedure. It is conceivable that the failure of the definitions to facilitate recognition in Experiment 1 would not have occurred had the more sensitive forced-choice procedure been used. (2) The cued-recall test of the first experiment (included to demonstrate a knowledge of list membership of at least some unrecognized items) was replaced by a cued-recognition test.

Method

Materials and design. Two variants of the critical list were obtained by modifying those of Experiment 1 in a way that did not allow subjects to score on the cued-recognition test on the basis of normative association between the cue and target terms. Half of the cues in each of these lists were dropped and each

Table 1
Variants of Critical List in Experiment 2

| | Target | |
Cue	Variant 1	Variant 2
ground	COLD	HIGH
head	LIGHT	STUPID
memory	NEED	SLOW
whiskey	GREEN	WATER
barn	LARGE	DIRTY
art	DAY	GIRL
pretty	BLUE	RED
cave	WET	BLACK
moth	BALL	FOOD
brave	WIND	WEAK
glue	CHAIR	WORK
country	MAN	OPEN
fruit	FLOWER	CAKE
cottage	SWEET	LOVE
grasp	BABY	HAND
roll	SMOOTH	RUG
drink	SMOKE	SLEEP
exist	PAIN	BEING
cloth	SHEEP	COAT
swift	GO	BIRD
lady	QUEEN	SHORT
glass	CUT	HARD
plant	BUG	TREE
cabbage	WASH	ROUND

remaining cue was paired with both its original target and one from the other variant. This yielded two variants of the critical list matched with respect to cue. These are shown in Table 1.

In addition to the critical list, two 24-pair set-establishing lists were used. These were the same as the set lists in Tulving and Thomson's (1973) experiments. The target definitions were the same as in Experiment 1.

The design was similar to that for Experiment 1. The procedure was the same for all subjects, except that half were given one variant of the critical list and half the other. The 24 target-lure pairs involved in the testing sequence were presented in random order. The definitions of the items within a test pair were read in the same order for all subjects; since a given test word was a target for half the subjects and a lure for the other half, the order of reading the target and lure definitions was balanced between subjects.

Subjects and procedure. The subjects were 12 young adults who were paid for participating. The procedural details up to and including presentation of the critical list were as in Experiment 1, except that the short set-establishing list of the first experiment was replaced by two full-length (24-pair) set lists. An interference task was interpolated between the presentation and testing of the critical list. This task consisted of the presentation and immediate free recall of nine 18-word lists; none of the words was a member of the critical list. The critical testing session involved the presentation of a list of word pairs, such as COLD-HIGH. One member of each pair was a target word, the other its yoked lure (i.e., the word which, for other subjects, was a target paired with the same list cue). The subject made three successive forced-choice recognition judgments for each pair of words, and for each judgment gave a confidence rating according to a 6-point scale. After making his first judgment, the subject read the dictionary entries for both test items and then made his second judgment. He was then given the relevant list cue (e.g., *ground*), whereupon he made his third judgment. He then followed the same procedure with the next pair of words, and so on until three judgments had been made on all 24 word pairs. No time limits were imposed.

Results

As in Experiment 1, the data fail to support the hypothesis that reading the definitions of the target words increases their probability of being recognized. Thus, with a total possible of 288, the number of targets selected in the first and second judgments was 216 and 218, respectively. Similarly, there was no net change in the recognition confidence ratings for the two judgments: A total of 48 ratings were revised in the target direction (more confidence for designated targets, less for lures), and 52 were revised in the lure direction (less confidence for targets, more for lures).

Analysis of the cued-recognition task, which followed the second judgments, indicated that a substantial number of the targets for which the dictionary definitions failed to prompt recognition as critical list items were in fact potentially recognizable. Thus, a total of 256 of the 288 target items were designated targets in the cued-recognition test. By applying high-threshold corrections, we may conclude that only 51% of the targets were recognized in the second judgment, though at least 78% of the targets were potentially recognizable.

Our conclusion that definition reading does not benefit recognition holds when the first and second judgments are compared only for those items designated targets in the cued-recognition test. In fact, such conditional analyses showed for both hit rate and mean confidence rating a very slight decline from the first to the second judgment.

DISCUSSION

The failure to increase the probability of recollecting the presentation of a word by providing its dictionary definition seems contrary to the modified version of tagging theory.

It could be argued that this evidence against the theory rests on the acceptance of the null hypothesis. However, this point loses much of its force when we compare (1) the total lack of success (and not merely a nonsignificant success) of prompting the target word with its dictionary entry, and (2) the very substantial success of prompting the target word with the single word with which it had been presented.

Of course, the hypothesis that cuing a target word with its dictionary entry should facilitate recognition is in some respects quite unreasonable. No doubt the subjects thought so; they averaged well over a minute for each definition and probably doubted that they were of any use. But the important point is that the hypothesis does seem to follow from modified tagging theory, which attributes recognition failure of (at least) recallable items to a failure of access to the appropriate sense nodes. Obviously some of the recognition failure in the present experiments could be attributed to the tagging of nodes representing idiosyncratic senses; nevertheless, the range of meanings covered in the dictionary is broad and should cover a substantial proportion of the nodes nested under the target words. And since the "understanding" a subject achieves on reading the successive senses of a dictionary presumably signals access to their corresponding nodes, it seems that the failure to increase recognition with the procedures described in this paper is contrary to modified tagging theory.

REFERENCES

MARTIN, E. Generation-recognition theory and the encoding specificity principle. *Psychological Review*, 1975, **82**, 150-153.

REDER, L. M., ANDERSON, J. R., & BJORK, R. A. A semantic interpretation of encoding specificity. *Journal of Experimental Psychology*, 1974, **102**, 648-656.

TULVING, E., & THOMSON, D. M. Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 1973, **80**, 352-373.

WATKINS, M. J., & TULVING, E. Episodic memory: When recognition fails. *Journal of Experimental Psychology: General*, 1975, **104**, 5-29.

Webster's new collegiate dictionary. Springfield, Mass: G. & C. Merriam, 1973.

NOTE

1. We should note that the failure to find an increase in recognition from the first to the second judgment cannot be attributed to nonspecific interference. Although the second judgment of a word was made after an average retention interval that was slightly greater than that for the first judgment, the pattern of results is unchanged when, for each subject, the first judgment of the first target word tested and the second judgment of the last one tested are eliminated, even though these omissions necessarily leave a mean delay that is slightly greater for the *first* judgment.